APPENDIX A

STATISTICAL CALCULATIONS

A.1 Data Set Characterization

A.1.1 Data Set Identification and Sorting

Statistical methods were used to evaluate and interpret the validated data generated during the Pit B Pre-design Study (Pit B PDS). The purpose of this analysis is to determine whether the Pit B waste contains a constituent in excess of its characteristically hazardous concentration. For this analysis, the data were pooled with those obtained during the Supplemental Site Investigation at Pit B conducted in November, 1995. For any analyte for which at least one positive detection exceeded its characteristically hazardous concentration, a 95% confidence upper confidence limit of the mean was established and compared to the applicable regulatory value.

A.1.2 Frequency of Detection Sorting

The statistical methods that were used to compute UCLs are largely dependent on the frequency of detection of a given analyte within a given data set. Given the fact that a non-detected result in and of itself implies some uncertainty of the concentration of a result between the sample-specific sample quantitation limit (SQL) and zero (0), the handling of non-detects for statistical purposes is paramount. Furthermore, as the proportion of non-detects in a data set increases, so does the uncertainty in the summary statistics computed on these types of data sets. For this reason, USEPA [1992] recommends different procedures for dealing with data sets containing certain ranges of non-detects. USEPA [1992] recommends the segregation, by proportion of non-detects, of data sets into four classes, each of which utilizes different methods to compute a valid UCL on the principal indicator of central tendency. These classes are listed below:

A. Data Sets Containing Between 0% and 15% Non-Detects
B. Data Sets Containing Between 15% and 50% Non-Detects
C. Data Sets Containing Between 50% and 90% Non-Detects
D. Data Sets Containing Between 90% and less than 100% Non-Detects

For the purposes of the Pit B PDS, GeoSyntec classified the data sets listed above as types A, B, C, and D, respectively. Statistical analysis procedures for each type of data set are described individually below.

A.2 Data Set Specific Statistical Analysis Procedures

A description of the types of summary statistics computed and statistical procedures to be used on each data set type are provided below by data set type under consideration.

A.2.1 Summary Statistics

As a component of the statistical analysis, several summary statistical parameters were computed for certain types of data sets. These summary statistics and the data set types to which they were applied are summarized below.

A.2.1.1 Mean

The mean is the average of the values in the data set. It was computed according to the formula:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where:

n = the number of points in the data set
x_i = an individual datum within the data set
$\bar{x}$ = the data set mean

Means were computed for Data Set Types A and B only.

A.2.1.2 Standard Deviation

The standard deviation of a data set is the square root of the mean squared deviations from the data set mean. It was computed according to the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{(n-1)}}$$

where:

s = the data set standard deviation
n = the number of points in the data set
x_i = an individual datum within the data set
$\bar{x}$ = the data set mean

Standard deviations were computed for Data Set Types A and B only.

A.2.1.3 Median

The median is the middle value (50% quantile) in a data set when the number of points in the data set is odd. It is computed as the midpoint between the two middle values (midpoint of the 50% quantile) when the number of points in the data set is even. The median was computed for all data set types (A through D).

A.2.2 Data Set Type A (0%-15% Non-Detects)

For data sets falling into this category, one-half of the sample-specific SQL was substituted for the non-detected values. The Shapiro-Wilk W test for normality [Shapiro and Wilk, 1965] was then conducted on the data set at 95% confidence (a=0.05), with one-half of the SQLs substituted for the non-detected values.

The Shapiro-Wilk W test is an effective test of whether the underlying distribution being tested is normally distributed. Data normality is a prerequisite to the computation of certain types of statistical intervals (e.g., parametric upper confidence limits (UCLs)). A discussion of this testing procedure follows. In the Shapiro-Wilk Test, the following hypothesis is tested [Gilbert, 1987]:

H_0 : The population has a normal distribution
H_1 : The population does not have a normal distribution

If H_0 is rejected, then H_1 is accepted and the population is concluded to not be normally distributed. If H_0 cannot be rejected, then there is no reason to doubt that the population is normally distributed, given the data set tested. To make this determination, a W test statistic was computed. The denominator, d, of this statistic was computed using the formula:

$$d = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

where:

n = the number of points in the data set
x_i = an individual datum within the data set
$\bar{x}$ = the data set mean

Then the data were ordered from largest to smallest to obtain sample order statistics. For example:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Then, k was computed by the following formula:

$$k = n/2 \text{ if } n \text{ is even}$$

$$k = (n-1)/2 \text{ if } n \text{ is odd}$$

The coefficients $a_1, a_2, a_3, \ldots, a_n$ were then determined from tabulated values provided in Gilbert [1987], and the W test statistic was computed by the formula:

$$W = \frac{1}{d}\left[\sum_{i=1}^{k} a_i \left(x_{[n-i+1]} - x_{[i]}\right)\right]^2$$

If the W statistic was less than the W quantile at a=0.05 (95% confidence) (provided in Gilbert, [1987]), or if the P value of the test was less than a (0.05), then H_0 was rejected, and the population was concluded to be not normal. If the W statistic exceeded the W quantile at a=0.05 (95% confidence), or if the P value of the test was greater than a (0.05), then H_0 was not rejected, and there was no reason to doubt the normality of the population. The P value of this test is the probability associated with the computed W statistic. If it is less than the significance level (a) selected for the test, this is an indicator that the null hypothesis should be rejected. If is not less than the significance level selected for the test, then this is an indicator that the null hypothesis is probably appropriate and should be retained.

If these data tested positive for being normally distributed, then a parametric UCL at 95% confidence was constructed on the data set according to the methods outlined by USEPA [1986, 1989, and 1992]. A parametric UCL is a statistical interval that is designed to estimate the mean of the population with a given level of confidence.

UCLs were computed by the following formula [USEPA, 1986, 1989, and 1992]:

$$UCL = \bar{x} + t_{0.05,(n-1)} \times \frac{s}{\sqrt{n}}$$

where:

$\bar{x}$ = the data set mean
s = the data set standard deviation
n = the number of points in the data set
$t_{0.05,(n-1)}$ = Student's t statistic at 95% confidence and $(n-1)$ degrees of freedom

If the data did not test positive for normality, then the natural logarithms of the data were computed, using one-half of the SQL as an input parameter. The SQLs were halved prior to logarithmic transformation. The natural logarithms of the data set (with the natural logarithm of one-half of the SQL substituted) were tested for normality using the Shapiro-Wilk W test at 95% confidence (a=0.05) as described above. If the natural logarithms of the data tested positive as being normally distributed, then a parametric UCL that achieved 95% confidence was constructed on the log-transformed data set as described above. In this case, the UCL of the logarithms was compared to the logarithm of the regulatory limit to determine whether a compound is present at a concentration exceeding the hazardous threshold. If the natural logarithms of the data did not test normal as described above, then a non-parametric UCL that achieved a confidence of 95% was placed on the median according to the general nonparametric method for estimating quantiles outlined in Gilbert [1987].

This method is outlined below. The data are first ordered from smallest to largest as follows. For this application, the SQL is used as the input parameter for these calculations.

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

If n>20, the one-sided upper limit is given by the following formula.

$$u = p(n+1) + Z_{(1-\alpha)}\sqrt{np(1-p)}$$

where:

p= the quantile (percentile) of interest; for the median, p=0.5
n= the number of points in the data set
Z= the normal statistic at (1-α)% confidence

The value computed for u represents the rank (order) of the value corresponding to the 95% UCL on the median. If, for example, u=13, the 95% UCL is the thirteenth value from the minimum, or x_{13}. If u is not an integer, linear interpolation can be used to determine the value for the UCL which lies between the two ordered statistics bounded by u.

A.2.3 Data Set Type B (15%-50% Non-Detects)

A.2.3.1 Normality Testing

To determine the normality of these types of data sets, Censored and Detects Only Probability Plots were constructed according to procedures outlined in USEPA [1992]. To construct the Censored Probability Plot, the combined set of detects and non-detects was ordered, with non-detects being assigned arbitrary, but distinct ranks. The ith ordered value of the sample was designated as x_i, and m_i represents the approximate expected value of the ith ordered normal quantile, calculated as follows:

$$m_i = \Phi^{-1}\left(\frac{i}{(n+1)}\right)$$

where Φ^{-1} is the inverse of the standard Normal distribution with zero mean and unit variance. The values for x_i were plotted on the x axis whereas the values for m_i were plotted on the y axis.

To construct the Detects Only Probability Plot, the non-detects were completely ignored and the detects only were ordered. Values for m_i and x_i were computed as described earlier and plotted as described for the Censored Probability Plot.

To ascertain which adjustment procedure should be used, the correlation coefficient, r, of both the Censored and Detects Only Probability Plots was computed by the formula provided by Ott (1984):

$$r = \frac{\sum_{i=1}^{n}(x_i)(m_i) - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} m_i\right]}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} m_i^2 - \frac{\left(\sum_{i=1}^{n} m_i\right)^2}{n}\right]}}$$

An equivalent formula for the computation of r is provided in USEPA [1992]. The absolute value of r for the probability plots of the original (non-transformed) data set were compared for both the Censored and Detects Only Probability Plots. The plot with the highest r is the more linear, and the adjustment technique (Cohen's or Aitchison's) associated with the more linear plot was used to adjust the mean and standard deviation. The correlation coefficient of the most linear plot was compared to tabulated values in USEPA [1992]; if it exceeded the appropriate critical value tabulated therein, then normality was concluded.

If a data set was concluded to be not normally distributed, then the natural logarithms of the data set were computed. The above testing regime was applied to them, and a similar comparison of correlation coefficients was made.

In the event that none the data were neither normally nor lognormally distributed, then a non-parametric UCL was placed on the median according the generalized method for estimation of quantiles presented by Gilbert [1987] and compared to the appropriate regulatory value.

A.2.3.2 Cohen's Adjustment Procedures

The construction of a Censored Probability Plot is a test of the underlying assumptions of Cohen's adjustment method, specifically that non-detects are real values that exist below the SQL and that have been censored at their detection limit. If the data plotted in the Censored Probability Plot were more linear than the Detects Only Probability Plot (i.e., if the absolute value of the Censored Probability Plot r value exceeded that of the Detects Only Probability Plot) and if the SQLs of the non-detected values did not vary, then Cohen's adjustment [Cohen, 1959 as reported in USEPA, 1989] was used to correct the mean and standard deviation of the original data set if the data set was normally distributed or of the log-transformed data if the data set was lognormally distributed.

If any of the criteria required for the use of Cohen's adjustment (as outlined in paragraph one of this subsection) were not satisfied, or if, based on best professional judgment, the

assumptions governing the use of Aitchison's adjustment were more germane to the data set under consideration, then the latter procedure (Aitchison's adjustment) was considered.

A.2.3.3 Aitchison's Adjustment Procedures

The Detects Only Probability Plot is a test of the underlying assumptions of Aitchison's adjustment, specifically that non-detects represent zero values and that non-detects and detects follow separate probability distributions. If the Detects Only Probability Plot was more linear (i.e., the absolute value of the correlation coefficient of it exceeded that of the Censored Probability Plot) than the Censored Probability Plot, or if the SQLs of the non-detected values vary, then Aitchison's adjustment [Aitchison, 1955 as reported in USEPA, 1992] was used to correct the mean and standard deviation of the original data set only. This procedure was not applied to log-transformed data sets because of a statistical anomaly in the adjustment procedure.

A.2.3.4 UCL Calculation

Regardless of the adjustment method employed, if the data set was normally distributed, a parametric UCL that obtained 95% confidence was constructed as described previously, using the adjusted mean and variance.

If the data set was neither normally nor lognormally distributed, or if the data set was normally or lognormally distributed but either adjustment could not be performed for reasons specific to each and described above, then a non-parametric UCL was placed on the median using the method outlined by Gilbert [1987] for general nonparametric estimation of quantiles. This value was compared to the regulatory limit to determine whether constituents were present at concentrations at levels exceeding hazardous levels.

A.2.4 Data Set Type C (50%-90% Non-Detects)

A non-parametric UCL at 95% confidence was placed on the data set median using the method outlined by Gilbert [1987] for general nonparametric estimation of quantiles without any *a priori* distribution testing. No testing was needed in this case because the proportion of non-detects was so large that any assumptions or adjustments used to compensate for them would be meaningless due to the large degree of uncertainty in these data sets.

A.3 Statistical Calculations

A.3.1 Benzene

SAM	VALUE	LOGS
A1	268	5.59099
A2	333	5.80814
A3	465	6.14204
A4	47.3	3.85651
AB1	15.8	2.76001
B1	277	5.62402
B2	200	5.29832
B3	25.4	3.23475
C1	222	5.40268
C2	445	6.09807
D1	<50	3.21888
D2	79.3	4.37324
D3	87.9	4.4762
E1	1350	7.20786
E2	1130	7.02997
F1	1910	7.55486
F2	1780	7.48437
F3	1440	7.2724
F4	25.4	3.23475
GPBW1	1800	7.49554
GPBW2	70	4.2485
GPBW3	150	5.01064
GPBW4	440	6.08677

There is one non-detect in this data set, corresponding to a percentage of non-detects of 4.3%. Therefore, the protocol for Data Set Type A was used to compute a 95% UCL.

Normality Test (Raw Data):

$W_{stat} = 0.759782$ $W_{crit}(95\%) = 0.914$

Reject H_0, conclude raw data are not normally distributed at 95% confidence.

Normality Test (Logarithms):

$W_{stat} = 0.937795$ W_{crit} (95%) = 0.914

Cannot reject H_0, no reason to doubt the normality of the log-transformed data at 95% confidence.

UCL Computation

The UCLs will be calculated based on the logged data.

$\bar{x}_{log} = 5.413456$

$s_{log} = 1.529266$

$n = 23$

$t_{0.05,22} = 1.717$

$$UCL = \bar{x}_{log} + t_{0.05,22} \frac{s_{log}}{\sqrt{n}}$$

$$UCL = 5.413456 + 1.717 \times \frac{1.529266}{\sqrt{23}} = 5.9609$$

The TCLP value for benzene is 500 µg/L. The natural logarithm of this value is 6.21.

The UCL for benzene is less than the logarithm of the TCLP value; therefore benzene is not present at hazardous concentrations in the Pit B waste.

A.3.2 1,2-Dichloroethane

SAM	VAL2	LOGS
A1	502	6.2186
A2	474	6.16121
A3	872	6.77079
A4	3.29	1.19089
AB1	1.68	0.51879
B1	440	6.08677
B2	165	5.10595
B3	<50	3.21888
C1	238	5.47227
C2	812	6.6995
D1	<50	3.21888
D2	22.1	3.09558
D3	<50	3.21888
E1	57.6	4.05352
E2	45.8	3.82428
F1	113	4.72739
F2	475	6.16331
F3	235	5.45959
F4	<50	3.21888
GPBW1	100	4.60517
GPBW2	<2500	7.1309
GPBW3	100	4.60517
GPBW4	420	6.04025

There are five non-detects out of a total of 23 points. This corresponds to a proportion of non-detects of 21.3%. Therefore, the data set will be handled according to the procedures outlined for Data Set Type B.

Normality Testing

Untransformed Data (Censored Probability Plot):

OBS	CODE	VALUE	NORMAL
1	N	502.00	0.96742
2	N	474.00	0.67449
3	N	872.00	-1.38299
4	N	3.29	-1.38299
5	N	1.68	-1.73166
6	N	440.00	0.54852
7	N	165.00	0.10463
8	N	238.00	0.31864
9	N	812.00	1.15035
10	N	22.10	-1.15035
11	N	57.60	-0.31864
12	N	45.80	-0.96742
13	N	113.00	-0.00000
14	N	475.00	0.81222
15	N	235.00	0.21043
16	N	100.00	-0.15753
17	N	100.00	-0.15753
18	N	420.00	0.43073

Note: "N" indicates a detected value

The correlation coefficient for the line fitted to this data is 0.8805; the critical value at 95% confidence is 0.945. Therefore, it can be concluded that the data are not normally distributed under the assumptions for Cohen's adjustment.

Untransformed Data (Detects Only Probability Plot)

OBS	CODE	VALUE	NORMAL
1	N	502.00	1.00315
2	N	474.00	0.63364
3	N	872.00	1.61986
4	N	3.29	-1.25212
5	N	1.68	-1.61986
6	N	440.00	0.47951
7	N	165.00	-0.06601
8	N	238.00	0.19920
9	N	812.00	1.25212
10	N	22.10	-1.00315
11	N	57.60	-0.63364
12	N	45.80	-0.80460
13	N	113.00	-0.19920
14	N	475.00	0.80460
15	N	235.00	0.06601
16	N	100.0	-0.40777
17	N	100.00	-0.40777
18	N	420.00	0.33604

Note: "N" indicates a detected value

The correlation coefficient for the line fitted to this data is 0.9312; the critical value at 95% confidence is 0.945. Therefore, it can be concluded that the data are not normally distributed under the assumptions for Aitchison's adjustment.

Log-Transformed Data (Censored Probability Plot)

OBS	CODE	VALUE	LOGS	NORMAL
1	N	502.00	6.21860	0.96742
2	N	474.00	6.16121	0.67449
3	N	872.00	6.77079	1.38299
4	N	3.29	1.19089	-1.38299
5	N	1.68	0.51879	-1.73166
6	N	440.00	6.08677	0.54852
7	N	165.00	5.10595	0.10463
8	N	238.00	5.47227	0.31864
9	N	812.00	6.69950	1.15035
10	N	22.10	3.09558	-1.15035
11	N	57.60	4.05352	-0.31864
12	N	45.80	3.82428	-0.96742
13	N	113.00	4.72739	-0.00000
14	N	475.00	6.16331	0.81222
15	N	235.00	5.45959	0.21043
16	N	100.00	4.60517	-0.15753
17	N	100.00	4.60517	-0.15753
18	N	420.00	6.04025	0.43073

Note: "N" indicates a detected value; "LOGS" are the natural logarithms of the values; NORMAL are the normalized logs.

The correlation coefficient for the line fitted to this data is 0.96; the critical value at 95% confidence is 0.945. Therefore, it can be concluded that the data are lognormally distributed under the assumptions for Cohen's adjustment.

Log-Transformed Data (Detects Only Probability Plot)

OBS	CODE	VALUE	LOGS	NORMAL
1	N	502.00	6.21860	1.00315
2	N	474.00	6.16121	0.63364
3	N	872.00	6.77079	1.61986
4	N	3.29	1.19089	-1.25212
5	N	1.68	0.51879	-1.61986
6	N	440.00	6.08677	0.47951
7	N	165.00	5.10595	-0.06601
8	N	238.00	5.47227	0.19920
9	N	812.00	6.69950	1.25212
10	N	22.10	3.09558	-1.00315
11	N	57.60	4.05352	-0.63364
12	N	45.80	3.82428	-0.80460
13	N	113.00	4.72739	-0.19920
14	N	475.00	6.16331	0.80460
15	N	235.00	5.45959	0.06601
16	N	100.00	4.60517	-0.40777
17	N	100.00	4.60517	-0.40777
18	N	420.00	6.04025	0.33604

Note: "N" indicates a detected value; "LOGS" are the natural logarithms of the values; NORMAL are the normalized logs.

The correlation coefficient for the line fitted to this data is 0.931; the critical value at 95% confidence is 0.945. Therefore, it can be concluded that the data are not lognormally distributed under the assumptions for Aitchison's adjustment.

Only one probability plot, the Censored Probability Plot of the log-transformed data, showed significant linearity such that a conclusion can be made as to normality of the values considered. The analysis indicates that the data are log-normally distributed under a set of assumptions (i.e., those required for Cohen's adjustment) and that Cohen's

adjustment should be applied. However, one of the requirements for the use of Cohen's adjustment is that the SQLs of the data set do not vary; for the above data set, there is variation in the SQLs of the data set, and, therefore, Cohen's adjustment cannot be applied, and a non-parametric UCL must be placed on the data set.

UCL Calculation

The non-parametric UCL will be computed by the below formula:

$$u = p(n+1) + Z_{(1-\alpha)}\sqrt{np(1-p)}$$

$$p = 0.5$$

$$n = 23$$

$$\alpha = 0.05$$

$$Z_{0.05} = 1.645$$

$$u = 0.5(23+1) + 1.645\sqrt{(23)(0.5)(0.5)}$$

$$u = 15.94$$

The true estimate of the median at 95% confidence lies 94% of the distance between the 15th and 16th ordered statistics. These values are 238 and 420, respectively. The distance between them, Δ, is 182. 94% of Δ is 171.08. Therefore, the 95% UCL of the median TCLP concentrations for 1,2-dichloroethane is 409.08 µg/L. The value for a determination of a characteristically hazardous waste with respect to 1,2-dichloroethane is 500 µg/L.

The UCL for 1,2-dichloroethane is less than the logarithm of the TCLP value; therefore 1,2- dichloroethane is not present at hazardous concentrations in the Pit B waste.

A.4 References

Aitchison, J. 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of American Statistical Association* 50: 901-908.

Cohen, A.C., Jr. 1959. Simplified estimators for the normal distribution when samples are single censored or truncated. *Technometrics* 1: 217-237.

Gilbert, 1987. *Statistical Methods For Environmental Pollution Monitoring.* New York, Van Nostrand Reinhold.

Ott, L. 1984. *An Introduction to Statistical Methods and Data Analysis.* Second Edition. Boston, Duxbury Press.

Shapiro, S.S. and Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52: 591-611.

USEPA, 1986 *Test Methods for Evaluating Solid Waste, SW 846, 3rd edition,* November, 1986.

USEPA, 1989. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities: Interim Final Guidance*, Office of Solid Waste, Washington, D.C., April, 1989.

USEPA, 1992. *Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance*, Office of Solid Waste, Washington, D.C., June, 1992.